The Gist of Everything New: Personalized Top-k Processing over Web 2.0 Streams

Sebastian Michel
Saarland University, Saarbrücken, Germany
smichel@mmci.uni-saarland.de

Joint work with Parisa Haghani and Karl Aberer, EPFL, Lausanne

27/10/2010
Outline

Motivation and Problem Statement

Profile Selection

Result Maintenance

Experiments
Outline

Motivation and Problem Statement

Profile Selection

Result Maintenance

Experiments
To bring you up to date......

The latest news......
To bring you up to date......

Police Taser 'Buck-Naked' Jogger (Video)

backporch.fanhouse.com — Eighteen-year-old Anthony King to go for a jog buck naked in broad daylight. And perhaps most puzzling of all: he was wearing swimming gog...
To bring you up to date......

Police Taser 'Buck-Naked' Jogger (Video)
backporch.fanhouse.com — Eighteen-year-old Anthony King went for a jog buck naked in broad daylight. And perhaps most puzzling of all: he was wearing swimming gog...  

Top Five Lame Halloween Costumes
blogs.houstonpress.com — We came up with five (and an extra) dos and don’ts for this Halloween as warnings for what to avoid at all costs.  

S. Michel The Gist of Everything New
To bring you up to date......

Police Taser 'Buck-Naked' Jogger (Video)
backporch.fanhouse.com — Eighteen-year-old Anthony King decided to go for a jog buck naked in broad daylight. And perhaps most puzzling of all: he was wearing swimming googles...

Top Five Lame Halloween Costumes
blogs.houstonpress.com — We came up with five (and an extra two) do's and don'ts for this Halloween as warnings for what to avoid at all costs.

OBEN OHNE

Warum haben Profis keine Brustbehaarung?

Fußball-Stars sind cool und sexy. Auf Oben-Ohne-Fotos fällt auf: Kaum ein Kicker hat Haare auf der Brust. Superstar Ronaldo rasiert sich überall. mehr...
To bring you up to date......

Sarkozy most unpopular French President in 50 years

According to the French poll, ahead of another week of strikes, petrol shortages and travel delays, the President's popularity has gone down to just 29 per cent, 3 points less than September, to reach its lowest ever.

more »
To bring you up to date......

**Sarkozy most unpopular French President in 50 years**

According to the French poll, ahead of another week of strikes, petrol shortages and travel delays, the President’s popularity has gone down to just 29 per cent, 3 points less then September, to reach its lowest ever.

*more »*

---

**The Drunk CookBook**

madatoms.com — The story of every college going kid who has fun while at it

*5 Comments  Save  47 min ago*
RT @cillycicilia: #NED, we ? you! Ciattt!!!
PingChat Wow! @PingChat! is giving away an iPhone4 here! http://bit.ly/aQHwbz
y’all #FF @starterbeauty and subscribe to her youtube #naturalhair
RT @iheartquotes: "A leader leads by example not by Force." ~Sun Tzu
#FF She swam the Atlantic to see me but got lost and ended up in Greenland fishing village @spiffykates
#panas ... Berapa kartu yang akan terbit malam ini ? #need #bra
is stuck with watching #bra vs #ned on CCTV5, which is in Chinese... i love it.. she does all the work for me ;) #lazy
Trailer de la mejor pelcula para Geeks: Java 4ever [Humor] [Video] http://dlvr.it/2F1Y5 #Video
Robinho estressou ali, hein... #MEDO! kkk
@diasemglobo #umdiasemglobo Sem Mas!
muscle... #WTF ...just thot you would all like to know this breaking news! LoLz
RT @H_FJ: If there were a holiday to celebrate kerning, it would be 7/4... #fontlaffs
RT @RT_Sitoiaussi: Si toi aussi tes amis c’est comme une deuxieme famille ? #RT_Sitoiaussi
Gracias Leni! :) RT @lenizahdez: #FF pa las chicas @TatianaPerez... Looks like a beautiful weekend ahead! Are you doing anything fo
Motivation

Users can not keep up with the new information

- Existing solution:
  - actively subscribe to different sources (RSS feed)
  - receive all updates or,
  - receive updates based on relevance (how to set a threshold?)

...Still the result is also a stream of data, usually from a few sources

Relevance?

- Only a small portion of the whole is interesting
- Relevance is *relative*: How interesting is a document for a user?
- Desire for *Recency*: Even most relevant items are too old after some time
Problem Statement

- Users specify their interest as profiles (large number of users)
  - A set value of term, weight
- Stream of incoming documents specified by:
  - ID (i.e. URL of a blog)
  - A set value of (term, weight)
  - Arrival time
- In-order streams: documents arrive in the same order as generated
- Temporal factor: documents are valid while belonging to a sliding window
Problem Statement

- Users specify their interest as profiles (large number of users)
  - A set value of term, weight
- Stream of incoming documents specified by:
  - ID (i.e. URL of a blog)
  - A set value of (term, weight)
  - Arrival time
- In-order streams: documents arrive in the same order as generated
- Temporal factor: documents are valid while belonging to a sliding window

Goal
Large scale personalized top-k processing over text streams in window.
Goal

For each user generated profile:

- **continuously** keep the top-k list of “active” documents
  - new documents come in (insert to top-k list?)
  - old documents expire (removed from top-k list?)

So? ... Why is this a problem?

- continuous changes on high rates
- which profiles are affected? all?
- expensive re-evaluation if top-k doc expires
Overall Structure

Profile Selector
- term-profile inverted index
  - new document
  - new document & affected profiles

Result Maintenance
- Profiles hash table
- Result set
- Top-k processing over valid docs
  - profiles
  - time sorted list
  - expired document
  - updated results
  - profiles to be re-evaluated

S. Michel The Gist of Everything New
Avoid processing all profiles against a new document.
Avoid processing all profiles against a new document.

Avoid re-evaluation each time a document expires.

Result Maintenance

Top-k processing over valid docs.
Outline

Motivation and Problem Statement

Profile Selection

Result Maintenance

Experiments
Profile Selection

Naive Solution

- avoid evaluating all profiles
- maintain a term-profile inverted index
  - for each term keep a list of profiles which contain t
- upon document arrival compare against all profiles in the document’s terms lists

\[ d = \{(\text{obama},0.4), (\text{chicago},0.2), \ldots (\text{basketball},0.1)\} \]
Profile Selection (COL)

Completely Ordered Lists

- keep sorted inverted lists
- enable early stopping
- for a profile $p$
  - $p.s$: score of ranked $k$ document for $p$
  - $u_i$: $t_i$’s weight for $p$
  
  $$v_i = \frac{u_i}{p.s}$$

Example

$(worldcup, 0.4), (southafrica, 0.3), (2010, 0.3)$

Top-k score = $0.20$

$$v_{worldcup} = \frac{0.4}{0.20} = 2$$

$$v_{southafrica} = \frac{0.3}{0.20} = 1.5$$

$$v_{2010} = \frac{0.3}{0.20} = 1.5$$
Profile Selection (COL)

Upon document arrival...

- round-robin processing of lists
- stopping condition:
  - Let $v_i$ be the last observed value in list $l_i$
  - Stop when $\sum v_i w_i < 1$
- Effectively avoids processing all profiles in a list!

Drawbacks

- the lists should be kept sorted all the time
- each time a profile is updated all its corresponding entries should be updated
- this cost can overshadow the benefits of early stopping!
Profile Selection (POL)

Partially Ordered Lists

- decrease the cost of maintaining sorted lists
- do not keep complete order
- entries ordered with regard to a number of boundaries

structure with **constant** insert and delete cost

...
Boundary Selection

- How to select effective boundaries?
- Minimize extra cost POL incurs with regard to COL
- Given a term t assume we know:
  - pdf of weight of t in documents: $f_W$
  - pdf of corresponding v values: $f_V$
- Intuition:
  - Finer granularity where $pr(V \cdot W < 1)$ is larger
Outline

Motivation and Problem Statement

Profile Selection

Result Maintenance

Experiments
The effect of sliding window...

- When a document which was a top-k result for p expires, p should be re-evaluated
- Re-evaluations can be very costly
- How to avoid re-evaluations all together
  - For each profile p keep all docs which have a chance to become a top-k
  - Keep the k-skyband!
Keeping only the necessary

![Graph showing points O_1, O_2, O_3, O_4, and O_5 on a graph with Score on the y-axis and Lifetime on the x-axis, connected by a line.]
Result Maintenance

In-order streams:

- **ALL** incoming docs are part of k-skyband of **ALL** profiles
- skyband changes with every income!
- all profiles should be updated!

Approach: Keep **only part of the skyband** = The Horizon

- above score threshold $\tau$
- more stable
- more promising
Result Maintenance

In-order streams:

- **ALL** incoming docs are part of k-skyband of **ALL** profiles
- skyband changes with every income!
- all profiles should be updated!

Approach: Keep **only part of the skyband** = The Horizon

- above score threshold $\tau$
- more stable
- more promising

How to choose $\tau$?

- static fixed value not suitable
- dynamic value adapting to the window
Outline

Motivation and Problem Statement

Profile Selection

Result Maintenance

Experiments
Experiments

Dataset

- ICWSM 2009 Spinn3r Blog Dataset
- 44 million blog posts during 2 months
- first tier blogs: 2 444 780
- Each blog post consists of plain text, a timestamp, a set of tags, and URL
- TF*IDF methodology used for assigning weights replayed according to the timestamps attached to the entries
- Profiles:
  - frequently used topic tags of blogs, i.e. US elections
  - 600 unique terms
  - each profile has between 3-5 terms
- similarity function: cosine
Methods under Comparison

State of the Art

- Mouratidis et al. [ICDE '09]: incremental threshold method
- for each profile remember the state of index list scan positions
- + documents seen during scan
- adapt scan lines accordingly when docs arrive/expire above these thresholds

Our Methods

- profile selection: COL (complete ordered list), POL (partial ordered)
- result maintenance: keep k, horizon

+ the naive implementation
Some Results

![Graph 1: Average Total Time vs. Window Size](image1)

- naive-horizon
- col-horizon
- pol-horizon
- incr.thresh.

![Graph 2: Average Total Time vs. Number of Profiles](image2)

- simple-horizon
- col-horizon
- pol-horizon
- incr.thresh.

S. Michel

The Gist of Everything New
Some Results

![Graph](image)

- **#profiles=50000 k=10 w=4500**
- **average total time (ms)**
- **number of groups**
- **#profiles=50000 k=10 w=4500**
- **average profile filtering time (ms)**
<table>
<thead>
<tr>
<th>Algorithms</th>
<th>total</th>
<th>profile fil.</th>
<th>update</th>
<th>re-eval</th>
<th>insert</th>
</tr>
</thead>
<tbody>
<tr>
<td>naive-k</td>
<td>23.12</td>
<td>7.45</td>
<td>0.00</td>
<td>14.73</td>
<td>0.67</td>
</tr>
<tr>
<td>col-k</td>
<td>23.87</td>
<td>2.36</td>
<td>3.12</td>
<td>17.39</td>
<td>0.74</td>
</tr>
<tr>
<td>pol-k</td>
<td>21.14</td>
<td>3.48</td>
<td>0.52</td>
<td>16.17</td>
<td>0.71</td>
</tr>
<tr>
<td>naive-horizon</td>
<td>11.01</td>
<td>7.18</td>
<td>0.00</td>
<td>1.71</td>
<td>1.80</td>
</tr>
<tr>
<td>col-horizon</td>
<td>9.64</td>
<td>3.47</td>
<td>1.30</td>
<td>2.71</td>
<td>1.83</td>
</tr>
<tr>
<td>pol-horizon</td>
<td>8.84</td>
<td>4.60</td>
<td>0.28</td>
<td>1.78</td>
<td>1.85</td>
</tr>
<tr>
<td>incr.-thresh</td>
<td>13.86</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Conclusion

- Efficient continuous top-k processing
- Need to check many user profiles against stream of documents
- Early stopping of profile list scans
- Skyline based result maintenance to avoid re-evaluations